CSE8803-EPI

Using machine learning model and statistics model to forecast different targets of Covid-19

Project Milestone

Te-Kai Chen
Computational Science and Engineering
Georgia Institute of Technology
Atlanta GA US
tchen483@gatech.edu

Chen Chen
College of Computing
Georgia Institute of Technology
Atlanta GA US
chesterchen@gatech.edu

1 Introduction

The Covid-19 pandemic has caused worldwide unrest and forcing local organization to measure to reduce its spread. In order to accurately forecast the impact of the disease, we will use machine learning model and statistical model to investigate on multiple targets of covid-19 data including cases, mortality, hospitalizations, and reproduction number (Rt). Our time range for dataset would be taken from January.21 2020 to October.25 2021. This project (Milestone) will focus on the main frame of machine learning based model and non-machine learning based model to estimate our targets. It will observe the target which is easier to forecast and evaluate the output to compare the advantage and disadvantage of each model. We will find out the best model that suitable for our dataset and indicate the performance. Furthermore, we will divide US into four different regions to find out the region which can be used to describe the whole US. We will also use the five most populated county in the four regions to construct a network forecasting our targets. For both machine learning model and non-machine learning model, this project will introduce the algorithm and proposed ideas to analyze data and try to find out the best model trained by regional data that can forecast properly for coronavirus outbreak in US.

2 Review of the relevant prior work

2.1 Machine Learning based models

We are going to use four machine-learning-based models to solve the forecasting problem. We use the time-series model to deal with this problem since our objectives are to estimate Covid-19 cases, mortalities, hospitalizations, and Rt, which are all time-related. In [9, 13], they introduced some data-driven machine learning methods that are used to perform the real-time forecasting. We then choose four models to deal with the problem. The first model is the standard Recurrent Neural Network (RNN) [1]. RNN is a model which mainly has the input of sequential data. The difference between RNN and the typical feed-forward network is that RNN will consider the current information and the previous inputs. Hence, RNN can perform better than the regular feed-forward network with this characteristic when time is a crucial factor in data. The second one is Echo State Network (ESN) [2]. The ESN model is a modification of RNN having a reservoir that randomly connects nodes inside the hidden layer. With this modification, ESN can solve some chaotic situations in the network. The third model is Long Short-Term Memory (LSTM) [14]. LSTM also has the same ideas as RNN but different from the inner blocks of the node. The LSTM block introduces the concept of gate preserving the portion of the information with current and previous information. Due to the ideas, it can help the model to learn whether the previous information is important or not. In [3], the model use LSTM to encode the data and use the attention layer to focus on the important sequence. The last model is Gated Recurrent Unit (GRU) [4]. GRU is a modified version of LSTM because it preserves the gate's idea but replaces the gates with an update gate and reset gate. This modification can help the network with fewer parameters and make the training faster and more stable. In [10], it proposed a model named MODELPREDENC using GRU as the neural network unit to encode history data to help forecast target of Covid-19. We will use the configuration in [11] to set up our models.

2.2 Non-Machine Learning based models

Previously, we chose several models for non-machining learning-based model and try to calibrate the parameters for each one. At the beginning of model selection, there are multiple options. In the meantime, in terms of evaluation metrics, the root mean square error (RMSE)[17] was employed to assess the performance of each time series model. We narrow down to select three models to address the relationship with Covid-19. After the proposal, we are going to compare its definition and mathematical logic between various models, including ARIMA[5], Hot-Winters Additive Model (HWAAS)[6], Prophet. In simple words, HWAAS is the model to extend from Holt's exponential smoothing, adds the seasonality factor to the forecast, which is best for data with trend that do not fluctuate or increase, and it is the choice of smoothing parameters and the normalization of indices. Secondly, Prophet[8] is the model which developed by Facebook to predict business-related problems and issues. Prophet can use a decomposable time series model, which it used to predict the Covid-19 cases. Other than that, we find out ARIMA is the most well-known and widely used families of time-series models includes the auto-regressive integrated moving average models. ARIMA assume a linear correlation between time series values and attempt to estimate the dependencies. In this milestone, we find out the models that used to estimate, and use RMSE (root mean square error) as an indicator to each of the model. Throughout the project milestone, we achieve the definition of different models and their underlying behavior. However, we haven't implemented the entire stimulation at this moment, which will continue to the finalize in the final project.

3 Data collection process

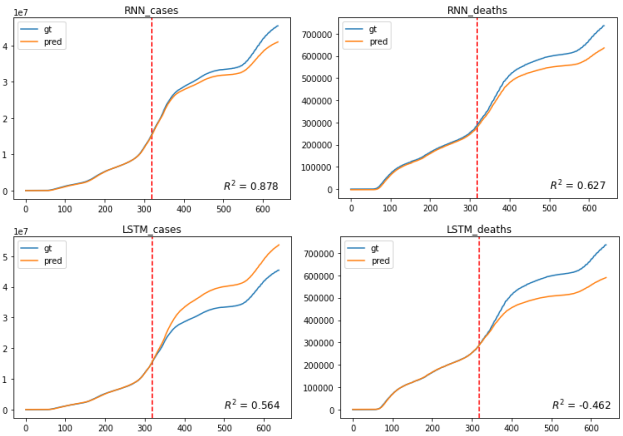
For the milestone project, we are tending to use several datasets to evaluate our model in the following circumstances. Since the dataset is recommended from CSE8803-EPI course piazza. The New York Times (NYTimes) is releasing data with cumulative counts of coronavirus cases in the United States, this not only provides us with another form of visualization, but it also uses a formal, statistical model underneath. We are going to choose NYT as our datasets, which contains the number of cases in various parts of the country. One of the advantages of NYT dataset is that the daily by using the whole dataset to predict our targets, we have the ability to understand the underlying meaning between each model. In other words, we can explore and find out the advantage of different models with these evaluations. In other words, the updated data could be helpful for us to evaluate the performance of the forecasting model. Datasets contains several files that are considered as EXCEL format such as us-counties-recent.csv, us-counties.csv, us-states.csv and us.csv. each excel file include various indexes including date, cases, deaths, states, counties.

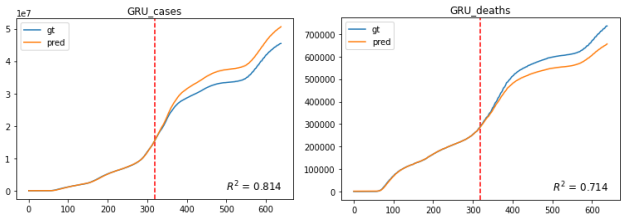
The Covid-19 dataset constructed by NY Times is the option that we select, since it is a daily updated dataset which it contains the most recent data for us to estimate the performance of the forecasting model more easily. Secondly, there are various excel files that containing covid-19 information in NY times dataset. We pick "us-states.csv", "us-counties.csv", "CDC Hospitalization Dataset by daily", "The total population (co-est2020" and "Rt_data" as history data as our target dataset in order to the next level of investigation.

4 Initial findings/Summary statistic of datasets

4.1 Initial findings

We have constructed the normal vanilla RNN, LSTM, GRU model to predict the cases and deaths with the "us.csv" dataset. The result for each model is presented in below diagram. We use the first 50 percent of data to train the model and forecasting rest of the data. We found out that RNN has the best performance on forecasting total cases and GRU has the best performance on deaths. However, LSTM doesn't perform very good, which has the lowest R^2 for both cases (0.564) and deaths (0.462).





4.2 Summary statistic of datasets

First, "us-states.csv" consist of date, state, fips, cases, deaths, there are 34k rows and date range is from 2020-01-21 to 2021 the most recent 2021-11-01. To explore the mortality, we can split to different region of states (north-east, mid-west, west, south). Which it would be districted by four different regions. To be more specially. The northeast are Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, and Pennsylvania. The mid-west are Ohio, Michigan, Indiana, Wisconsin, Illinois, Minnesota, Iowa, Missouri, North Dakota, South Dakota, Nebraska, and Kansas. The south are Delaware, Maryland, Virginia, West Virginia, Kentucky, North Carolina, South Carolina, Tennessee, Georgia, Florida, Alabama, Mississippi, Arkansas, Louisiana, Texas, and Oklahoma. The west are Montana, Idaho, Wyoming, Colorado, New Mexico, Arizona, Utah, Nevada, California, Oregon, Washington, Alaska, and Hawaii. Initially, we group by state as four different regions above and sum up deaths to estimate the trend of death cases.

The second dataset we considering is "us-counties.csv", which is similar to "us-state.csv". It contains date, county, state, fips, cases and deaths. At this moment, we are interested in the top n counties with the most population that we could analyze on the death cases trend by top n population counties. It is beneficial to merge population dataset and us-counties dataset to find the top n population of counties.

Besides, we use "Rt.data" as history data and the total population as "co-est2020.csv" as the total population to estimate the mortality. In addition, we use the CDC Hospitalization dataset by daily to evaluate the hospitalization of "impatient_beds_used_covid" as indicator, and estimate the trend related to the top n states. For example, we could visualize the relationship between top n states and impatient_bed_used during covid-19.

5 Mathematical backgrounds for problem

To address the mathematical background related to this problem, our metrics will be accuracy MAE (median absolute error), RMSE (Root-Mean-Square-deviation), and R^2 (coefficient of determination) to evaluate our statistical models and minimize the error to figure out the best performance. For example, we could use the Root-Mean-Square-deviation (RMSE) to collect each time series model and compare different approaches in different chiromancers. Furthermore, we will indicate the results corresponding to the performance of models and find out which model would be the better option for estimate the Covid-19. For instance, RMSE will be applied to predictive results and approaches, that it is significant for us to instantiate the model for certain situations by comparing methods for different datasets. To calculate the Root-Mean-Square-Error (RMSE), we could calculate the residual for each data point, compute the norm of residual for each date point, and then compute the mean of residuals and take the square root of that mean, where N is the number of data point. In machining learning, RMSE is extremely helpful to have a single number to judge model's performance, which is one of the most widely used measures for predication.

6 Description of algorithms used

For the description of algorithms it used, we used RMSE as metrics to evaluate NYT datasets, contributing with time series model HWAAS, ARIMA and Prophet. The RMSE is a measure of how far from the regression line data points are, which is a measure to concentrate the data is around the line of best fit. ARIMA provides a simple powerful method for making skillful time series forecasts. HWAAS is an exponential smoothing with additive trend and additive seasonality formulation. Prophet is the measure with autoregressive recurrent networks.

7 General difficulties during elaboration

There are multiple difficulties that we suffer during the project milestone. For building the machine learning model, the result shows that the normal vanilla RNN performs better than LSTM and GRU on both cases and mortality. We are not sure whether it's correct since the LSTM and GRU should perform better in general. We

think there are two reasons for this, the first one is that the cases and deaths don't rely much on the data which are far from current timestamp. The second one is that the hyperparameters for LSTM and GRU are not optimal. Since it's hard to find the optimal hyperparameter, we aren't sure what will be the optimal result of these models.

For building non-machine learning model, we want to use AtsPy library, which provides a way of automating the process of Time Series Forecasting. Moreover, Atspy contains a variety of Models such as HWAAS, ARIMA, Prophet, etc. It is a straightforward automating library and easy to use. However, while we want to import this library, there are multiple error/exceptions being thrown. The error comes from the initial package import, which is "from atspy import AutomatedModel", the error is similar to "No module named 'gluonts.trainer'". However, it still occurs after I import gluonts. Looking through stack Overflow and google, but still haven't figure it out the reason.

8 Tasks implemented later

There are several tasks will be implemented after the project milestone. For non-machine learning model, it focuses on AtsPy library, since it is an automating built in package to simulate time series models such as ARIMA (Automated ARIMA Modelling), Prophet (Modelling Multiple Seasonality With Linear or Non-linear Growth) and HWAAS (Exponential Smoothing With Additive Trend and Additive Seasonality). Solved the import error/exceptions while it suffers in the project milestone. Afterwards, to compare these models with its performance and accuracy related to the Covid-19 datasets.

For the part of machine learning, we will implement the remaining ESN model and try to tune all the model to have the best performance. After tunning the model, we will then try to Also, there are a lot of papers using attention to help, we may try to add the attention layer to find out whether it will help improve the performance.

Since we now only forecast the data of cases and deaths in "us.csv", we will later forecast other data with the corresponding data. However, since the Rt and hospitalization only have data of each state, we will only forecast these targets by states.

Last but not the least, we'll try to use the models of regions and five counties with highest population density to observe the one having the ability to forecast the whole us community.

9 Respond to the comments

- We add a title on the top of our report, which is "Using machine learning model and statistics model to forecast different targets of Covid-19"
- We have added new refs on section 2
- Mentioned cases/mortality/hospitalization/Rt in section 1
- We are now only using the past data to construct/train our model
- We are not going to use 'us-prisons.csv' and 'us-colleges.csv' as our dataset because they only contain 2020 and 2021 data with single value for each year.
- Time period of dataset would be taken from January.21 2020 to October.25 2021.
- We are tending to use Rt and hospitalization as part of our datasets. It contains state information in both Rt and hospitalization.

REFERENCES

[1] H. T. Siegelmann and E. D. Sontag, “Turing computability with neural nets,” Applied Mathematics Letters, vol. 4, no. 6, pp. 77–80, 1991.

[2] Jaeger H, Haas H. Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. Science. 2004 Apr 2;304(5667):78-80.

[3] Yudistira N. COVID-19 growth prediction using multivariate long short term memory. arXiv. 2020;14(8):1–8.

[4] Chung, Junyoung, Gulcehre, Caglar, Cho, KyungHyun, and Bengio, Yoshua. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.

[5] Jamal Fattah, L. E. (n.d.). Forecasting of demand using ARIMA model - Jamal Fattah, Latifa Ezzine, Zineb Aman, Haj El Moussami, Abdeslam Lachhab, 2018.

[6] Yar, M., & Chatfield, C. (2002, April 23). Prediction intervals for the Holt-Winters forecasting procedure.

[7] Jebb, A. T., Tay, L., Wang, W., & Huang, Q. (2015, June 09). Time series analysis for psychological research: Examining and forecasting change.

[8] Taylor, B. S., & Letham, B. (2021, March 01). Prophet: Forecasting at scale.

[9] Siva R Venna, Amirhossein Tavanaei, Raju N Gottumukkala, Vijay V Raghavan, Anthony Maida, and Stephen Nichols. 2017. A novel data-driven model for real-time influenza forecasting. bioRxiv (2017), 185512

[10] Kamarthi, H., Rodríguez, A., & Prakash, B. A. (2021). Back2Future: Leveraging Backfill Dynamics for Improving Real-time Predictions in Future. arXiv preprint arXiv:2106.04420.

[11] Wang, Lijing, et al. "Examining Deep Learning Models with Multiple Data Sources for COVID-19 Forecasting." 2020 IEEE International Conference on Big Data (Big Data). IEEE, 2020.

[13] Hewamalage H., Bergmeir C., Bandara K. Recurrent Neural Networks for Time Series Forecasting: Current Status and Future Directions. Int.J. Forecast. 37, 388–427. (2021).

[14] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] Bijaya Adhikari, Xinfeng Xu, Naren Ramakrishnan, and B Aditya Prakash. 2019. Epideep: Exploiting embeddings for epidemic forecasting. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 577–586

[16] Snow, D. AtsPy: Automated Time Series Forecasting in Python. Atspy Automating Learning. (2020, April 20).

[17] Chai, Tianfeng. Root mean square error (RMSE) or mean absolute error (MAE)– Arguments against avoiding RMSE in the literature VL, 7 DO, 10.5194/gmd-7-1247-2014 JO, Geoscientific Model Development ER, AU, Draxler, R.R. PY, 2014

[18] .J.Mukherjee, M. Best research papers based on Facebook Prophet - Nerd For Tech. Time Series. Prophet Model (2021, August 11).